Conversation Level Constraints on Pedophile Detection in Chat Rooms

PAN 2012 — Sexual Predator Identification

Claudia Peersman, Frederik Vaassen,
Vincent Van Asch and Walter Daelemans

CLiPS
COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER

Universiteit Antwerpen

Task 1: Sexual Predator Identification

Preprocessing of the Data

- Data: PAN 2012 competition training set
  - info on the conversation, user and post level
  - predator ID list
- Two splits: training and validation set
- No user was simultaneously present in training and validation

→ prevent overfitting of user-specific features

Experimental Setup

- Features: token unigrams
- LiBSVM
- Probability output
- Parameter optimization
- Experiments on 3 levels
- data resampling

Level 1: the Post Classifier

- Resample the number of posts

→ Equal distribution of posts per class

- About 40,000 posts per class in training
- No resampling in the validation sets

Level 1: the Post Classifier (2)

- Only output on the post level
- Aggregate the post level predictions to the user level:
  - LiBSVM's probability outputs
  - Predators = average of the 10 highest predator class probabilities ≥ 0.85

Results for the Predator Class

Scores	Post Classifier
Recall	0.93
Precision	0.36
F-score	0.52

Level 2: the User Classifier

- Resampling on the user level

→ exclude users with no suspicious posts

- Filter: dictionary of grooming vocabulary
  → see Task 2
- Why?
  - reduce the amount of data
  - “hard” classification → higher precision?

Update Results (1)

Data reduction: up to 48.4%

Scores	Post Classifier	User Classifier
Recall	0.93	0.82
Precision	0.36	0.88
F-score	0.52	0.84

→ Combine systems?

Combining the systems

- Weighted voting using LiBSVM's probability outputs
- 70% of the weight on the high precision User Classifier

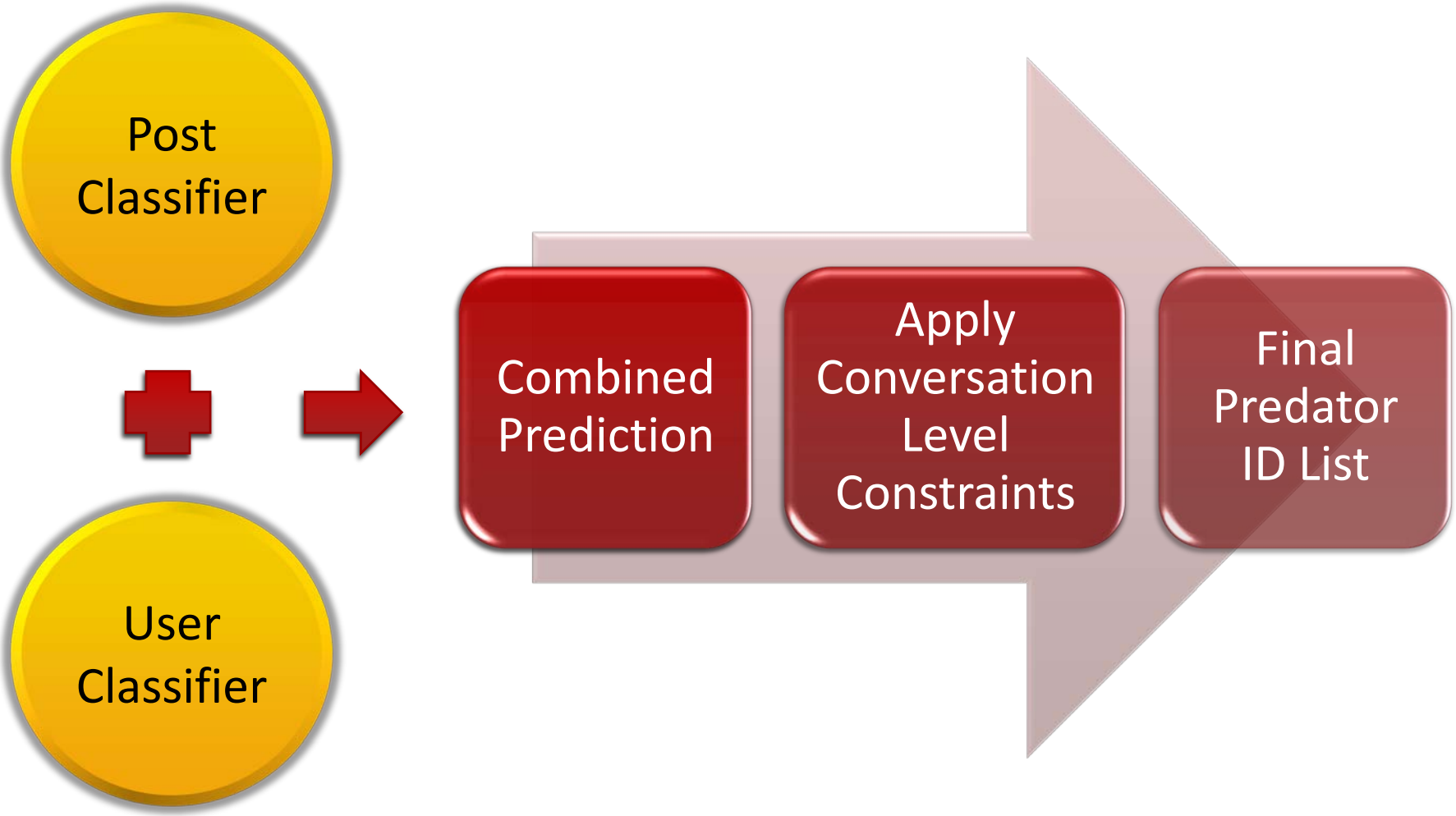
Update Results (2)

Scores	Post Classifier	User Classifier	Combined Results
Recall	0.93	0.82	0.85
Precision	0.36	0.88	0.84
F-score	0.52	0.84	0.84

Level 3: Conversation Level Constraints

- Both users in a conversation labeled as predators
- Our approach:
  - go back to predator probability output
  - use the high precision user classifier
  - Predator probability ≥ 0.75

System Overview



Update Results (3)

Scores	Post Classifier	User Classifier	Combined Results	Combined + Constraints
Recall	0.93	0.82	0.85	0.85
Precision	0.36	0.88	0.84	0.94
F-score	0.52	0.84	0.84	0.89

Results on the PAN 2012 Test Set

Scores	Combined + Constraints	PAN Test Set
Recall	0.85	0.60
Precision	0.94	0.89
F-score (β = 1)	0.89	0.72

- Future research:
  - more splits
  - investigate ensembles

Task 2: Identifying Grooming Posts

Identifying Grooming Posts

- From the final predator ID list → detect posts expressing typical grooming behavior
- No gold standard labels → What is grooming?
- Predator conversations have predictable stages (e.g. Lanning, 2010; McGhee et al., 2011)

Identifying Grooming Posts (2)

- Dictionary containing references to 6 stages:
  - sexual topic
  - reframing
  - approach
  - data requests
  - isolation from adult supervision
  - age (difference)

Identifying Grooming Posts (3)

- Resources:
  - McGhee et al. (2011)
  - English Urban Dictionary website http://www.urbandictionary.com/
  - English Synonyms http://www.synonym.net/
- cf. user classifier filter

Results on the PAN 2012 Test Set

- Precision = 0.36
- Recall = 0.26
- F-score (β = 1) = 0.30

Discussion

- Use of β-factors to calculate the F-score:
  - Task 1: focus on precision (β = 0.5)
  - Task 2: focus on recall (β = 3.0)
- However, in practice:
  - find all predators (recall in Task 1)
  - find the most striking posts (precision in Task 2)

Questions?

Contact: claudia.peersman@ua.ac.be

CLiPS
COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER

Universiteit Antwerpen